Open Research Online

The Open University's repository of research publications
and other research outputs

Propagating Data Policies: a User Study

Conference or Workshop Item

How to cite:

Daga, Enrico; d'Aquin, Mathieu and Motta, Enrico (2017). Propagating Data Policies: a User Study. In: Proceedings of the Knowledge Capture Conference, ACM, New York, NY, USA, article no. 3.

For guidance on citations see FAQs.

© 2017 The Authors

Version: Accepted Manuscript

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1145/3148011.3148022

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Propagating Data Policies: a User Study

Enrico Daga
The Open University
Milton Keynes, United Kingdom
enrico.daga@open.ac.uk

Mathieu d'Aquin
Insight Centre - NUI
Galway, Ireland
mathieu.daquin@insight-centre.org

Enrico Motta
The Open University
Milton Keynes, United Kingdom
enrico.motta@open.ac.uk

ABSTRACT

When publishing data, data licences are used to specify the actions that are permitted or prohibited, and the duties that target data consumers must comply with. However, in complex environments such as a smart city data portal, multiple data sources are constantly being combined, processed and redistributed. In such a scenario, deciding which policies apply to the output of a process based on the licences attached to its input data is a difficult, knowledge-intensive task. In this paper, we evaluate how automatic reasoning upon semantic representations of policies and of data flows could support decision making on policy propagation. We report on the results of a user study designed to assess both the accuracy and the utility of such a policy-propagation tool, in comparison to a manual approach.

CCS CONCEPTS

• **Information systems** → **Expert systems**; • **Computing methodologies** → **Knowledge representation and reasoning**; • **Applied computing** → *Law*;

KEYWORDS

Data Licences, Data Flows, Policy Propagation, User Study

ACM Reference Format:
Enrico Daga, Mathieu d'Aquin, and Enrico Motta. 2017. Propagating Data Policies: a User Study. In *Proceedings of Knowledge Capture, Austin, Texas USA, (Submitted, 2017) (K-CAP 2017),* 8 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In large data infrastructures such as City Data Hubs [8, 10], datasets are published under specific terms and conditions, which specify the actions that are permitted or prohibited, and the duties that target users must comply with (for example, the duty to attribute the data publisher). However, while datasets can be browsed in their original form, more often they are tailored to meet the needs of applications through complex manipulation processes that have new datasets as output. Such processing and republication of derived datasets can happen multiple times. In this scenario, a complex, knowledge intensive task is assessing what policies associated to the input of a process must be taken into account when deciding the terms and conditions applicable to its output [6]. For example, in a process producing aggregated statistics from two different input datasets, should the duty to attribute the source of the data or restrictions on commercial applications still apply to the result? In a previous work we developed a Policy Propagation Reasoner (PP Reasoner) to support data managers and users in making decisions on the policies associated with the generated datasets, on the basis of the ones associated with the input. By means of a semantic representation of policies and of data flows, the system combines OWL reasoning and Policy Propagation Rules (PPR) to compute the policies associated with any data node involved in the process [5]. In this work we report on a user study carried out to evaluate the system in terms of accuracy and utility to support the task of policy propagation as performed by data managers, processors and publishers. The participants were confronted with the problem of deciding what policies need to be taken into account when using a dataset that was derived from a complex process reusing licensed data sources. Their decisions were then compared with the ones of our system, and insights into its expected behaviour were acquired as well as observations about its accuracy and utility. In the next section we present the background and related work. In Section 3 we illustrate the PP Reasoner and the knowledge bases it relies upon. Section 4 describes the setup of the user study, the methodological criterias, and how the required resources were acquired and developed. We discuss the feedback received from the participants of the user study in Section 5, before going into the details of the results - Section 6 and discussing them - Section 7. Section 8 summarizes the contributions of this study and gives directions for future work.

2 CONTEXT AND RELATED WORK

Recently, numerous initiatives have investigated the vision of a *Smart City*, where cutting-edge technologies are applied to a number of sectors include government services, transport and traffic management, water, healthcare, energy, urban agriculture, waste, and resources management. City Data Hubs are emerging on the WWW as centralized nodes to control and monitor the flows of information between the variety of systems deployed in a given city or region [10, 11]. Current research aims to understand how to govern the life cycle of data in City Data Hubs [8]. The diversity of data sources, owners and licences associated with the data opens a new challenge, namely the problem of data exploitability, defined as the assessment of the policies associated with the data resulting from the computation of diverse datasets implemented within a City Data Hub [4]. Indeed, assessing how the policies associated with the sensed data will be propagated to the results of a data processing pipeline is an important problem. Data consumers might need to check which original sources of the data need to be acknowledged because of an attribution requirement, and even

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
K-CAP 2017, (Submitted, 2017), Austin, Texas USA
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$Priceless
https://doi.org/10.1145/nnnnnnn.nnnnnnn

whether the form of exposure or re-distribution they employ is allowed according to the policies attached to each individual piece of data they might obtain from the Data Hub. Research on policy models and reasoning focuses on the problem of licence compatibility and composition [2, 9]. However, reasoning on policy propagation is a necessary preliminary step to any policy validation or consistency check. In our work we reuse models developed within the Open Digital Rights Language[1] (ODRL) research community (for example [12]). A discussion on ODRL action dependencies and how they affect the policy semantics is included in [13]. Nonetheless, to the best of our knowledge, the first attempt to analyse how policies can propagate in manipulation processes is the one presented in one of our earlier papers [6]. In [6] we introduced the notion of

Policy	duty Attribution		
Output	lift-pareto-chart-1		
From	Property	To	
Food-Establishmen ...	isPartOf	input-data	
OpenDataCommuniti ...	isPartOf	input-data	
input-data	duplicate	input-data-1	
input-data	duplicate	input-data-4	
input-data-1	hasInterpretation	model-1	
input-data-1	hasComputation	model-1	
input-data-4	combinedIn	chart-data-1	
model-1	combinedIn	chart-data-1	
chart-data-1	remodelledTo	lift-pareto-chart-1	

Figure 1: Explanation: propagation trace.

Policy Propagation Rule (PPR) in order to solve the task of automatically deciding what policies associated to a data source need to be enforced to the output of a process in which that data source is involved. PPRs establish a fundamental connection between a policy - a permission, prohibition or duty - and a semantic relation between two data objects, expressed with the Datanode ontology [7]. Thus, it is possible to derive that a certain policy of the source needs to be enforced on the target [5]. The Datanode Ontology [7] allows us to model a data manipulation scenario as a network of data objects, making it possible to reason upon the relations between those data objects and to apply PPRs. However, in our earlier work we focused on the feasibility of the approach in terms of knowledge acquisition and management [6], scalability of the reasoner [5], and applicability in an end-to-end user scenario [4]. In this paper, we go a step further by performing a user experiment in order to evaluate the feasibility of policy propagation as a solvable problem and the hypotheses behind the development of the system, by relying both on quantitative and qualitative data analysis methods, particularly the Grounded Theory (GT) approach[2], in a comparison between the automatic process and a manual one performed by people with the typical skillet found in data consumers, processors and publishers who would be carrying out this task in a realistic context.

3 THE SYSTEM AT A GLANCE

The role of the PP Reasoner is to support users in the assessment of the impact of input data policies on the exploitation of the output data of processes and workflows. Consider the case where Food rating data released by a trusted authority under a licence that prohibits distribution is used alongside public data about city roads in order to assess the best Machine Learning approach, among several options, to employ for the prediction of good quality restaurants. This task would produce two types of outputs: (a) a set of datasets about roads labelled with the expected food quality rating; and (b) a set of datasets including details about the performance of each one of the algorithms tested. While the prohibition of distribution should be taken into account when using the former datasets, the same constraint would not apply to the latter.

The system is designed to work with a set of reference knowledge bases:

- *Data Catalogue.* Provides Datasets general metadata, including the link to the associated policy set (licence, Terms and Conditions, and so forth).
- *Licence Catalogue.* Includes the set of licences represented using the ODRL Ontology[3].
- *Process Catalogue.* Defines the set of processes represented using the Datanode Ontology[4].
- *Policy Propagation Rules (PPRs).* A rule base, developed and managed as described in [6]. Rules have the form of a connection between an atomic policy and a relation that is supposed to propagate it. For instance, `propagates(dn:cleanedInto, odrl:permission cc:DerivativeWorks)` instructs the reasoner to propagate `odrl:permission cc:DerivativeWorks` whenever a data item is `dn:cleanedInto` another, so that the cleaned item would also have the given policy.

The system was developed using the OWL reasoner of Apache Jena[5] in conjunction with a SPIN[6] rule engine. By relying on ODRL policies, Datanode graphs, and PPRs, the system computes the propagated policies for each node in the process graph[7]. The resulting RDF graph can be queried to obtain the policies of the output datanode. Moreover, the system can generate an explanation of the decision like the one in Figure 1, that traces the lineage of a given policy and highlights the arcs that would propagate it or block it[8].

[1] ODRL W3C Community Group: https://www.w3.org/community/odrl/

[2] https://en.wikipedia.org/wiki/Grounded_theory

[3] ODRL Version 2.1 Ontology: http://www.w3.org/ns/odrl/2/ODRL21

[4] Datanode Ontology: http://purl.org/datanode/ns/

[5] Apache Jena: http://jena.apache.org

[6] SPIN: http://spinrdf.org/

[7] More details about the implementation of the reasoner can be found in [4–6].

[8] The system's objective is to compute the set of propagated policies and it does not check the consistency of the policies. This could be done for the output set by relying on state of the art deontic reasoners like the one used in [9]. In the present work we are only interested in computing policy propagation in relation to the actions performed in different processes.

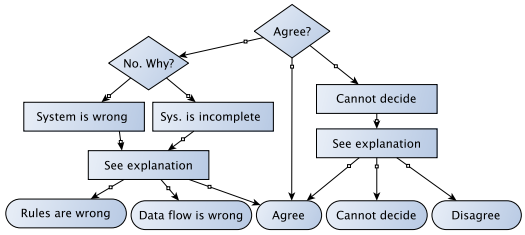

Figure 2: The decisions of the participants are compared with the ones of the system.

4 METHODOLOGY

The objective of the present study is to evaluate to what extend it is possible to support users on taking decisions upon the propagation of policies, and whether it is useful using the system outlined above. We assess this in two ways: 1) by comparing the decisions of the system with the ones performed manually in a quantitative analysis of the system's *accuracy*, and 2) by discussing the issues raised in disagreements with a qualitative analysis of the users' decision process. We further assess the value to users of the automatic support by asking the participants a set of questions concerning the experience of reasoning upon the policies and how they relate to processes, through a feedback questionnaire. In this section we illustrate the methodology employed in the study, including the *design* of the experiment, the criteria followed in the *sampling* of the users and the scenarios, and for data *collection* and *analysis*.

Design. The experiment simulated a set of scenarios in which a Data Hub manager needs to take a decision about which policies need to be associated to a dataset derived from a complex data manipulation process performed by the Data Hub infrastructure. We provided the participants with the same knowledge as the one used by the system, asking them to perform a number of decisions about policy propagation in reference scenarios, that we called *data journeys*. On each data journey, input datasets and the associated licences were presented, as well as a formalised representation of the process. Users were asked to take decisions about which ones of the policies derivable from the input licences should also be applied to the output data. We asked them only to decide whether a policy would propagate to the output, ignoring whether the process itself would violate the policies, or whether the propagated policies would be consistent with each other. Users were especially asked to compare their choices with the system, and discuss potential disagreements. The study was conducted with the support of a Web tool we developed, which guided the participants in the process.

A session started with and introductory phase, where the participants were given a short presentation about the Data Hub and the task they were going to perform, exemplified by a tutorial data journey. Then, the participants were left to face two *data journeys* involving real data. At the end of the sessions, users completed a feedback questionnaire individually.

A single data journey was structured as follows:

(1) **Understand the process.** Participants were asked to become familiar with the data process, described with the Rapid Miner tool[9].
(2) **Understand the input datasets.** In this phase the tool listed the dataset(s) selected to be the input source(s) and, for each dataset, the set of permissions, prohibitions or duties associated[10].
(3) **Indicate what policies shall propagate to the output.** Users are asked to indicate whether each one of them should be applied to the output, in the form of likert questions: *(-2) Certainly not*, *(-1) Probably not*, *(0) I don't know*, *(+1) Probably yes*, or *(+2) Certainly yes*.
(4) **Compare with the automatic reasoner.**This phase is summarised in Figure 2. The choices of the users are compared with the ones of the system, and conflicting ones are highlighted[13]. The journey terminated after all disagreements were discussed.

In all cases, a *propagation trace* was proposed to the users, as explanation of the system's decision (see Figure 1). Users could either: a) agree with the representation, but indicate that some relations should behave differently (for example, that `dn:hasCopy` should propagate the duty of `cc:Attribution`); b) disagree on the way the data flow was represented, and indicate how it should be; c) change their opinion after seeing the explanation and agree with the system. In cases where users could not decide, they were asked to justify why they believed they could not decide. We also gave them the possibility to abort the task, showing why a decision could not be made. In all cases, we asked them to compare their decision with the one of the system and to examine the explanation (*propagation trace*, see Figure 1), in order to collect insights into what to fix in the system.

Sampling: participants and scenarios. Ten participants were involved, selected among researchers and PhD students in our university, all having a background that includes some data analytics skills. The absence of a specific legal expertise in the study participants is intended. We evaluate the system with users who would typically perform such tasks. These people are developers, data scientists and practitioners who would process, reuse and republish data. Realistically we cannot assume them to have legal knowledge. To improve the quality of the decisions, we grouped the participants in teams of two persons, asking them to develop an agreement before taking action. Moreover, we introduced one intentional anomaly in the system, to check that users were paying enough attention during the study. We will refer to the five teams as follows: *MAPI*, *ILAN*, *CAAN*, *ALPA*, and *NIFR*.
Much effort was allocated to setting up realistic scenarios, comprising real data sources used in conjunction with real processes. The MK:Smart project has collected a large quantity of data sources about the city of Milton Keynes [8]. The datasets used in the study

[9]In particular, they were suggested to answer the following questions: what is the nature of the input data? What is the purpose of the process? What are the intermediary steps of the process? What is the nature of the ouput data?
[10]Participants were asked to check their understanding of the nature of the actions that are mentioned (also documented by the tool according to their specification in the related semantic web ontology - often ODRL but also CC[11], LDR[12]).
[13]It is worth noting that, while the users were requested to express an opinion with some degree of uncertainty, the system would always return a boolean answer: the policy is propagated or it is not.

are real data sources selected from the MK Data Hub data catalogue[14]. In order to select realistic workflows to be used in our experiment, we searched for pre-existing processes, instead of designing ad-hoc resources. Rapid Miner[15] is a popular tool that supports users in the design of articulated processes by means of a graphical user interface, making it a good candidate for the selection of our exemplary processes. Therefore, we explored the open source projects available on GitHub[16] searching for Rapid Miner process files. We selected *five* workflows representative of common data intensive tasks that could be applied to MK:Smart data. We designed five scenarios by associating them with real datasets from the MK Data Hub. From these associations, the *data journeys* listed in Table 1 were designed. Each data journey has some exemplary characteristic. *SCRAPE* refer to the very common expedient of crawling data out of web resources in order to setup a textual corpus. The *FOOD* journey is the scenario already mentioned in Section 3, where a data source is used to evaluate a Machine Learning approach. *CLOUD* refers to the extraction of textual data from micro blogs. There are many kind of statistical operations that can be performed on data, *AVG* is about the calculation of a *moving average*. A large part of the effort of applications relying on sensors is put in data preparation. This aspect is well reflected in the *CLEAN* data journey.

Each team was given 2 data journeys, and each data journey two teams. Later, we will compare the system twice on each scenario, and the teams between each other, in the agreement analysis. The data journeys were allocated following a latin square, thus avoiding to assign two tasks to the same two groups. We chose scenarios that were (a) complex enough, (b) feasible within 2 hours (so people would not be too fatigued), and (c) diverse enough in terms of use case (and type of operations performed). Although we cannot and do not formally claim for those scenarios to be a representative set of cases (because we cannot have all the cases), we can safely assume that they are ecologically valid. Also, the licenses are different in each scenario, covering a diverse range of policies: 15 permissions, 17 prohibition and 8 duties, selected from the 119 policies in the system. Overall, the experiment concerned 77 decisions including 40 policies, as often a policy was present in more than one scenario.

Table 1: Data Journeys.

SCRAPE	**Milton Keynes Websites Scraper.**
Websites about Milton Keynes are scraped, and indexed locally.	
Datasets	Milton Keynes Council Website (UK OGL 2.0), MK50 Website (All rights reserved), Wikipedia pages about Milton Keynes (CC-By-SA 3.0)
Policies	Permissions: reutilization, Reproduction, Distribution, DerivativeWorks. Prohibitions: Distribution, IPRRight, Reproduction, DerivativeWorks, databaseRight, reutilization, extraction, CommercialUse. Duties: Notice, ShareAlike, Attribution.
Process	https://github.com/mtrebi/SentimentAnalyzer/tree/master/process/scraper.rmp
Teams	*ILAN*, *MAPI*
FOOD	**Models for Food Rating Prediction.**
Two Machine Learning techniques are compared. The process uses data about Food Ratings and statistics about quality of life in MK wards and generates a lift chart and performance vectors.	
Datasets	OpenDataCommunities Worthwhile 2011-2012 Average Rating (UK OGL 2.0), Food Establishments Info and Ratings (Terms of use)
Policies	Permissions: DerivativeWorks, Distribution, Reproduction, display, extraction, reutilization. Prohibitions: modify, use. Duties: Attribution, Notice, display.
Process	https://github.com/samwar/tree/master/rapid_miner_training/16_lift_chart.rmp
Teams	*NIFR*, *ALPA*
CLOUD	**A tag cloud from microblog posts.**
Twitter posts about Milton Keynes are collected and processed in order to obtain a clean vector of words, associated with an occurrence score.	
Datasets	Twitter Feed #miltonkeynes (Terms of use)
Policies	Permissions: copy, display. Prohibitions: give, license, sell, transfer. Duties: attribute.
Process	https://github.com/jccgit/RM-Textmining-Pubmed/tree/master/Pubmed.rmp
Teams	*CAAN*, *ALPA*
AVG	**Moving average of sensors' records.**
Calculation of a moving average and plotting from sensor records.	
Dataset	Samsung Sensor Data (Terms of use)
Policies	Permissions: aggregate, anonymize, archive, derive, index, read, use. Prohibitions: CommercialUse, distribute, give, grantUse, move, sell, transfer. Duties: anonymize.
Process	https://github.com/billcary/Rapid_Miner/tree/master/chapter03/MovingAveragePlotter.rmp
Teams	*NIFR*, *MAPI*
CLEAN	**Sensor data cleaning workflow.**
The process performs a number of cleaning operations on sensors streams linked with postcodes in order to obtain a dataset ready for analysis.	
Datasets	Postcode Locations (UK OGL 2.0), Netatmo Weather Station - 52.022166, -0.806386, Netatmo Weather Station - 52.002429770568, -0.79804807820062 (Terms of use)
Policies	Permissions: CommercialUse, DerivativeWorks, Distribution, Reproduction, display, extraction, reutilization, use. Prohibition: Distribution, give, grantUse, license, transfer. Duties: Attribution, Notice, inform, obtainConsent.
Process	https://github.com/MartinSchmitzDo/RapidMinerDataCleaner/processes/clean.rmp
Teams	*CAAN*, *ILAN*

Data collection. During the experiment, we acquired three types of data: a) the decisions taken by the teams and the system, registered by the tool; b) a record of the motivations behind the decisions, in particular about disagreements with the system and borderline cases; and c) a feedback about the general difficulty of the task and the perceived user value of our system, obtained through a questionnaire including nine closed-ended leading questions and one single-choice question (see Table 2 and Figure 3). One of the authors also attended the study as supervisor providing support in the overall process - for example when the users needed clarification on the semantics of workflow actions or on the usage of the tool, but avoiding to influence their opinion on whether a policy ought to be propagated or not. Sessions' audio were recorded as well as the operations performed on the screen, preserving the discussions and small talks occurred motivating the rationales behind users' decisions.

Data analysis. We performed two different types of analysis: (a) an agreement analysis, to quantitatively measure the accuracy of the system; (b) a disagreements analysis, focused on discussing the quantitative results in the light of the users' justifications about controversial and borderline decisions, in a qualitative way. To assess the value to users of the support for policy propagation, we aggregated and discussed the responses of the questionnaire, that we present in Section 5. From the point of view of evaluating the accuracy and utility of the system, the quantitative data collected by the tool was expected to produce one of the following general results: a) teams agree with the system (and between each other), therefore the system is accurate; b) teams agree with each other that the system is not correct, therefore the task is feasible but the system need to be improved; or c) users don't agree with each other,

[14]MK Data Hub: http://datahub.mksmart.org
[15]Rapid Miner: https://rapidminer.com/
[16]GitHub: https://github.com/

and therefore the task of supporting automatically such decision is not feasible[17] The accuracy analysis is reported in Section 6, and complemented with a discussion on its statistical significance. A qualitative analysis was conducted by focusing on the disagreements and borderline cases selected from the quantitative results. To this aim, we transcribed the notes and conversations occurred during the experiment from the audio recordings and the tool. From these data we derived a set of general themes about fundamental issues on policy propagation, adopting a method that is akin to Grounded Theory (GT). We illustrate some exemplary cases and present the extracted themes in the discussion Section 7.

5 USER'S FEEDBACK

Before analysing the data journeys and how the decisions of the users relate to the behaviour of our system it is worth showing the feedback received after the study was conducted, collected through a questionnaire. In the questionnaire, we posed some questions about the problem of policy propagation to assess the value of the system to the user. The questionnaire was completed by the study participants individually. Table 2 summarises the nine closed-ended likert questions ($Q.1 - 9$), while Figure 3 shows the result of the single-choice question ($Q.10$). The majority of the participants of

Table 2: User's feedback. The shading of the cells reflect the distribution of the answers.

Q.ID	Question / *Left answ.*	<<	<	*Unsure*	>	>>	*Right answ.*
Q.1	How difficult was it to take a single decision on whether a policy propagates to the output?						
	Easy	0	1	3	6	0	Difficult
Q.2	Do you think you had enough information to decide?						
	Yes	2	6	2	0	0	No
Q.3	How difficult was it to reach an agreement?						
	Easy	1	5	2	2	0	Difficult
Q.4	Somebody with strong technical skills is absolutely required to take this decision. Do you agree?						
	Yes	1	8	1	0	0	No
Q.5	Somebody with strong technical skills is absolutely required to take this decision even with the support of automated reasoning. Do you agree?						
	Yes	1	5	1	3		No
Q.6	Understanding the details of the process is fundamental to take a decision. Do you agree?						
	Yes	6	3	1	0	0	No
Q.7	How enjoyable was it to discuss and decide on policies and how they propagate in a process?						
	Very	4	5	0	1	0	Not
Q.8	How feasible/sustainable do you think it is to discuss and decide on policies and how they propagate in a process?						
	Feasible	3	1	3	1	2	Unfeasible
Q.9	How sustainable do you think it is to discuss and decide on policies and how they propagate in a process with the support of our system?						
	Feasible	5	4	0	1	0	Unfeasible

our study believe that the task can be a difficult one ($Q.1$). However, the knowledge provided was adequate for making an informed decision ($Q.2$). Deciding whether a policy propagates is possible, even if not always trivial ($Q.3$). Users agree on considering policy propagation a problem that cannot be solved without understanding the details of the data manipulation process ($Q.6$), therefore someone with strong technical skills needs to be involved ($Q.4$, $Q.5$). The objective of $Q.7$ was to check whether users were positively invoved in the study, assuming that an unengaged person would

[17] In this last case, in fact, we would not be able to assess the accuracy of the system, and this might be evidence that the task cannot be solved at all, or at least that the knowledge bases used by the system are not sufficient to reason on policy propagation.

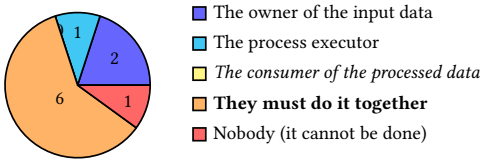

Figure 3: Q10. Who should decide on what policies propagate to the output of a process?

Table 3: Data Journeys: System decisions

Journey	*Propagated*	*Permissions*	*Prohibitions*	*Duties*
SCRAPE	15/16	4/5	8/8	3/3
FOOD	8/22	0/12	4/4	4/6
CLOUD	5/7	0/2	4/4	1/1
AVG	8/15	0/7	7/7	1/1
CLEAN	9/17	0/8	5/5	4/4
Tot	39/77	4/34	28/28	13/15

not put enough effort on expressing his opinion and taking thorough decisions. Questions $Q.8$ focused on the sustainability of the task. Users feedback on this matter was spread. Our hypothesis is that two data journeys are probably not enough to understand how much this task could scale in a real setting. However, our system can effectively support the user on taking a decision ($Q.1$, $Q.9$). This feedback shows that policy propagation is a difficult problem, although it can be solved with the right knowledge models. Therefore, a tool supporting this task has good value for users. The last question ($Q.10$) was meant to understand whether the Data Hub manager could actually decide on policy propagation. It turns out that most of the users think he/she cannot solve the issue alone, but he/she should involve the data owner and the process executor in this task. This conclusion reflects some of the issues raised during the study, that are discussed in Section 7.

6 ACCURACY ANALYSIS

In this Section we show how the decisions made by the users compare to the system. The decisions taken by the system are summarized in Table 3. For example, the *SCRAPE* data journey required to check 16 policies and the system decided to propagate 15 of them: 4 of the 5 permissions and all the prohibitions and duties. Tables 4a-4h summarize the results of our study in a quantitative way. The values are shown in two sets including the full numbers and the computed ratio, considering all the decisions (Tables 4a and 4b), and then split in *permissions* (Tables 4c and 4d), *prohibitions* (Tables 4e and 4f), and *duties* (Tables 4g and 4h). The values are first shown for each one of the user study (data journey of each team), aggregated for each data journey (average of both teams) and then as totals considering the decisions from all data journeys (at the bottom). The data journeys required from seven to twenty-two policies to be analysed for a total of seventy-seven *decisions*. Table 4a shows the number of decisions for each data journey (column D) and how much the teams agreed with the system (T_{avg} being the average value of the teams on the same data journey).

The agreement with the system is good, distributed differently across the data journeys and the teams, with an average ratio of 0.8. Moreover, this result is supported by the high agreement rate

Table 4: Agreement analysis.
D: total number of decisions; T_1, T_2: agreement between system and each team; T_{avg}: average agreement between teams and system; T_{12}: agreement between teams; T_{12+} agreement between teams (only *Certainly Yes*/*Absolutely No* answers); T_1+, T_2+: amount of *Certainly Yes*/*Absolutely No* answers. Tables on the left indicate totals, while the ones on the right show the related ratios.

(a) All decisions (totals)

Journey	D	T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
SCRAPE	16	15	13	14	14	11	11	14
FOOD	22	14	18	16	14	8	20	12
CLOUD	7	5	7	6	5	1	5	2
AVG	15	15	8	11.5	8	8	15	15
CLEAN	17	12	9	10.5	14	3	13	6
All	77			58	55	31	56.5	

(b) (ratios)

T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
0.9	0.8	0.9	0.9	0.8	0.7	0.9
0.6	0.8	0.7	0.6	0.6	0.9	0.5
0.7	1	0.9	0.7	0.2	0.7	0.3
1	0.5	0.8	0.5	1	1	1
0.7	0.5	0.6	0.8	0.2	0.8	0.4
		0.8	0.7	0.6	0.7	

(c) Permissions (totals)

Journey	D	T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
SCRAPE	5	4	2	3	3	0	0	3
FOOD	12	12	12	12	12	6	12	6
CLOUD	2	0	2	1	0	0	0	1
AVG	7	7	0	3.5	0	0	7	7
CLEAN	8	3	0	1.5	5	2	4	5
All	34			21	20	8	22.5	

(d) (ratios)

T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
0.8	0.4	0.6	0.6	0	0	0.6
1	1	1	1	0.5	1	0.5
0	1	0.5	0	N.A.	0	0.5
1	0	0.5	0	N.A.	1	1
0.4	0	0.2	0.6	0.4	0.5	0.6
		0.6	0.6	0.4	0.7	

(e) Prohibitions (totals)

Journey	D	T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
SCRAPE	8	8	8	8	8	8	8	8
FOOD	4	0	2	1	2	2	4	2
CLOUD	4	4	4	4	4	0	4	0
AVG	7	7	7	7	7	7	7	7
CLEAN	5	5	5	5	5	0	5	0
All	28			25	26	17	22.5	

(f) (ratios)

T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
1	1	1	1	1	1	1
0	0.5	0.3	0.5	1	1	0.5
1	1	1	1	0	1	0
1	1	1	1	1	1	1
1	1	1	1	0	1	0
		0.9	0.9	0.7	0.8	

(g) Duties (totals)

Journey	D	T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
SCRAPE	3	3	3	3	3	3	3	3
FOOD	6	2	4	3	0	0	4	4
CLOUD	1	1	1	1	1	1	1	1
AVG	1	1	1	1	1	1	1	1
CLEAN	4	4	4	4	4	1	4	1
All	15			12	9	6	11.5	

(h) (ratios)

T_1	T_2	T_{avg}	T_{12}	T_{12+}	T_{1+}	T_{2+}
1	1	1	1	1	1	1
0.3	0.7	0.5	0	N.A.	0.7	0.7
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	0.3	1	0.3
		0.8	0.6	0.7	0.8	

between the two teams (T_{avg} = 0.7). We observe that in more than half of the cases the decisions were made with the same degree of confidence (T_{12+}=0.6), and that in 70% of the cases users made a sharp decision about whether a policy would propagate or not (T_{1+}/T_{2+} total average is 0.6). Inspecting the table we see that the data journeys showing a lower agreement are $FOOD/T_1$, AVG/T_2 and $CLEAN/T_2$. We will discuss these in the next Section. The low scores on $CLOUD/T_{12+}$ and $CLOUD/T_{2+}$ only show a difference in the degree of confidence of the decisions, that is not especially relevant in this global view, altough this aspect will be discussed when looking at specific classes of policies.

Tables 4c and 4d only show results involving policies of type *permission*. The average agreement between the system and the users considering all the decisions is 0.6. Particularly, the *SCRAPE* data journey for T_2 shows a low agreement (0.6), also reflected in the number of common sharp decisions (0.4). This is a low score compared with the agreement ratio of *prohibitions* (0.9) and *duties* (0.8) that can be observed in Tables 4f and 4h. It is sufficient to consider at this stage how it was much easier to take decisions on prohibitions and duties, while permissions where a greater source of discussions and disagreements with the system. Moreover, decisions about prohibitions and duties appeared to be sharper than the ones about permissions, as both the agreement between the teams (T_{12}) and the choices with strong confidence (T_{1+},T_{2+}) received higher scores. However, on both types of policies the source of disagreement is on the *FOOD* data journey. We showed that this is the case (80% agreement). We complement this data with a statistical analysis based on the CohenâĂŹs kappa coefficient (CKC), that takes into account the possibility of the agreement occurring by chance. The 95% Confidence interval (CI) of CKC between the system and either human teams T1 or T2, is not significantly different from the 95% CI of CKC between T1 and T2. In other words, the system behaves as a user would do, also from a statistical point of view.

7 DISCUSSION

The results show that the task is feasible in all the scenarios and that our system exhibits good accuracy. In what follows we analyse the cases in which users disagreed with the system (also highlighted in Tables 4a-4h).

We expected three types of disagreements: a) the system is missing a policy; b) the system should block a policy; and c) the system should not decide about it as it does not have enough information. We note that option (c) never emerged from the study. The teams always made a clear decision whether to propagate a policy or not[18].

The *SCRAPE* data journey. Both teams agreed that the permission to `lds:extract` must be propagated. The system explanation showed that the policy was actually blocked by `dn:hasExtraction`. Both teams indicated this as an error, identifying the anomaly that was intentionally introduced, reassuring us about the committment in performing the task. $MAPI/T_2$ disagreed about propagating two specific permissions: `cc:Distribution` and `ldr:reutilization`. Although the general activity was one of web site crawling and indexing, $MAPI/T_2$ considered the type of indexing implemented to affect the interpretation of the content of the web site, in such a way to potentially damage the interests of the original owner: *"The process has a step in which some values were changed and then these changed values are assigned to be the LABELS of the items. [...] The permission to distribute of the new output should not be given for granted, to protect the owner of the content. This choice changes depending on the content of the data and not on the general action performed"*[19].

The *FOOD* data journey. This process produced two outputs, a *performance dataset*, including performance vectors of the machine learning algorithms compared, and a consequent *lift chart*, i.e. a graphical representation of the data. The system did not propagate any of the permissions (this decision aligns with the two teams). However, $NIFR/T_1$ changed their decision *after* seeing the explanation given by the system. Moreover, $NIFR/T_1$ decided that no policy should propagate to the output of the process, while $ALPA/T_2$ agreed with the system that both prohibitions and duties must be preserved for the performance dataset output, and only

[18] Although one participant observed in the questionnaire that the task of deciding on the policies to apply to a derived dataset was impossible.
[19] What is relevant here is not the point itself, that is arguable, but the fact that the participants believed the process and the policies were not enough to decide whether to propagate the policy.

the duties in the case of the *lift chart*. This difference is motivated by the interpretation of the nature of the *performance dataset*. By analysing the conversation occurred in the user study, it emerges that $NIFR/T_1$ considered the dataset containing only measures of the performance of the algorithms, while $ALPA/T_2$ interpreted it as a labelled dataset, therefore containing an enhanced version of the input data. The correct interpretation is the one of $NIFR/T_1$, once this is reflected in the data flow, the system will block all the policies as they are not applicable to the *performance dataset*.

The *CLOUD* **data journey.** $CAAN/T_1$ disagreed on the behavior of a set of rules about two permissions: `odrl:copy` and `odrl:display`, and marked as wrong the behavior of: `dn:processedInto`, `dn:cleanedInto`, `dn:refactoredInto`, `dn:isPortionOf`, and `dn:combinedIn`. In particular: *"combinedIn should propagate because both of the inputs have the permission to copy, in case one of the two has not, it shouldn't. You need to reason on the combination to decide the propagation."* During the session, the team proposed to propagate the permissions to the combined node as soon as no prohibition is present.

The *AVG* **data journey.** $NIFR/T_1$ changed their mind about permissions after seeing the explanation of the reasoner. Since the dataset was *modified* it made sense that the permissions were not propagated. However, $MAPI/T_2$ disagreed with both the system and the other team, and identified the problem by inspecting the explanation of the system: relation `dn:remodelledTo` must propagate the various permissions: *"As far as we understood there is no bias introduced in the data, therefore the permission should be kept intact. The outcome of the process depends entirely on the input, without additional information. It's just a mathematical process that keeps the information intact"*. In fact, `dn:remodelledTo` is defined as "Remodelling refer to the translation of the data to another symbolic structure (model), while keeping the same interpretation". This is a case where the teams disagreed about the system, however the justification of $MAPI/T_2$ seems robust enough to accept the change of behavior of the `dn:remodelledTo` relation.

The *CLEAN* **data journey.** Both teams observed that `dn:combinedIn` should propagate the permissions involved while the system decided not to. The main argument was that the relation should consider the policies of all datasets involved in the combination, however, without knowing them, the system should propagate them and leave the decision to a consistency check to be applied at the later stage. In another discussion, it was observed how in some cases there is a dependency between policies. It is the case of *duties*, that are always in the context of a permission, therefore by propagating the former, the system should also propagate the latter. For example, the permission to use should be propagated as a dependency of the duty to obtain consent.

The issues illustrated can be grouped under the following general themes:

a) *Incomplete knowledge.* The knowledge base used by the system is not complete: rules should be added or modified in order to *fix* the behaviour with respect to certain policies and relations, using the methodology presented in [6]. Data flows should be accurate and include all the relevant relations.

b) *Data reverse engineering.* A recurrent theme for assessing whether permissions should propagate was the contingency of *data reverse engineering*, defined in software engineering as "the use of structured techniques to reconstitute the data assets of an existing system" [1]. We observe that the correct interpretation of the nature of the output is crucial. Therefore, it is of fundamental importance that the data flow description is accurate, including assessing how much the information of the input data source can be extracted from the output data. In some cases, the implemented data flow was not complete enough to reflect this issue.

c) *Content-dependent decisions.* It was argued that in one case the impact of the process on the output policies could not be assessed without inspecting the content of the data. We cannot argue against this in principle. However, we assume that new relations could be developed within the methodology of [6] in order to capture fine grained implications of process actions on policy propagation, making this a case of incomplete knowledge base.

d) *Dependant policies.* The approach of the system was to consider the policies in isolation, and focusing on their interaction with process actions. However, it is clear that policies on their own incorporate a number of dependencies, some of them derived from the semantics of the action involved (for example `odrl:copy` is a kind of `odrl:use`), others from the way they are formalised in policy documents (in ODRL, a duty is always declared in the context of a permission). See also [13] for a discussion on this. However, by knowing that a policy needs to be taken into account on the output of a given process, dependant policies can be extracted from the original policy document.

e) *The Legal Knowledge.* A general observation that many of the participants made is that this is a *legal* issue, therefore a legal expert should be involved in the definition of policy actions, process descriptors, and PPRs. On one hand, this suggests the importance of providing support to Data Hub managers on deciding on policy propagation, as we cannot expect this type of users to have legal knowledge. On the other hand, this highlights a more general issue. In fact, a validation of the system by legal experts would assume a legal framework covering the status of metadata-oriented automatic reasoners in the Rule of Law, which is currently missing [3].

8 CONCLUSIONS

In this work we evaluated an approach and a system to support the assessment of the policies propagating from a data source to a derived dataset in a Data Hub. Participants agreed that it is possible to decide whether a policy associated with a dataset needs to be associated with another derived dataset. The results of our user study demonstrated that the task can be solved automatically with a good degree of accuracy. By considering both the results of the user study and the feedback collected, the system is overall accurate and is of good value to users. The study also let emerge a set of critical aspects involved. It is important that the knowledge bases are complete, in particular that the process description does not hide any of the elements that could influence the propagation of policies, for example making it clear how much of the data of the input can be extracted from the output.

From this study, we can conclude that there is evidence of a fundamental correspondence between the possible kinds of data-to-data relations and the way they affect policy propagation. However,

more research is required in order to include in the knowledge base other aspects involved in policy reasoning. The rights of other stakeholders should be involved in the process, including the ones of the process executor (what action adds value to the information?), or the rights of the entities represented in the data (from businesses to private citizens).

ACKNOWLEDGMENTS

The authors would like to thank Francesco Osborne and Paul Mulholland for the constructive feedback on a previous version of this paper.

REFERENCES

[1] Peter H Aiken. 1996. *Data reverse engineering: slaying the legacy dragon.* McGraw-Hill Companies.

[2] Pierfrancesco Bellini, Marco Mesiti, Paolo Nesi, and Paolo Perlasca. 2017. Protection and composition of crossmedia content in collaborative environments. *Multimedia Tools and Applications* (2017).

[3] Pompeu Casanovas. 2015. Conceptualisation of rights and meta-rule of law for the web of data. (2015).

[4] Enrico Daga, Alessandro Adamou, Mathieu d'Aquin, and Enrico Motta. 2016. Addressing Exploitability of Smart City Data. In *International Smart Cities Conference (ISC2).* IEEE.

[5] Enrico Daga, Alessandro Adamou, Mathieu d'Aquin, and Enrico Motta. 2017. Reasoning with Data Flows and Policy Propagation Rules. *Semantic Web Journal* (2017).

[6] Enrico Daga, Mathieu d'Aquin, Aldo Gangemi, and Enrico Motta. 2015. Propagation of Policies in Rich Data Flows. In *Proceedings of the 8th International Conference on Knowledge Capture.* ACM.

[7] Enrico Daga, Mathieu dâĂŹAquin, Aldo Gangemi, and Enrico Motta. 2014. Describing Semantic Web Applications Through Relations Between Data Nodes. (2014).

[8] Mathieu d'Aquin, John Davies, and Enrico Motta. 2015. Smart Cities' Data: Challenges and Opportunities for Semantic Technologies. *IEEE Internet Computing* 19, 6 (November 2015), 66–70. http://oro.open.ac.uk/45429/

[9] Guido Governatori, Antonino Rotolo, Serena Villata, and Fabien Gandon. 2013. One License to Compose Them All. In *The Semantic Web–ISWC 2013.* Springer, 151–166.

[10] Rodger Lea and Michael Blackstock. 2014. City Hub: A cloud-based IoT Platform for Smart Cities. In *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on.* IEEE, 799–804.

[11] Edward A Lee, Björn Hartmann, John Kubiatowicz, Tajana Simunic Rosing, John Wawrzynek, David Wessel, Jan M Rabaey, Kris Pister, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. 2014. The Swarm at the Edge of the Cloud. *IEEE Design & Test* 31, 3 (2014), 8–20.

[12] Víctor Rodríguez-Doncel, Serena Villata, and Asunción Gómez-Pérez. 2014. A dataset of RDF licenses. In *Legal Knowledge and Information Systems. JURIX 2014: The Twenty-Seventh Annual Conference.*, Rinke Hoekstra (Ed.). IOS Press. https://doi.org/10.3233/978-1-61499-468-8-187

[13] Simon Steyskal and Axel Polleres. 2015. Towards Formal Semantics for ODRL Policies. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web.* Springer, 360–375.